Data Engineering Fundamentals

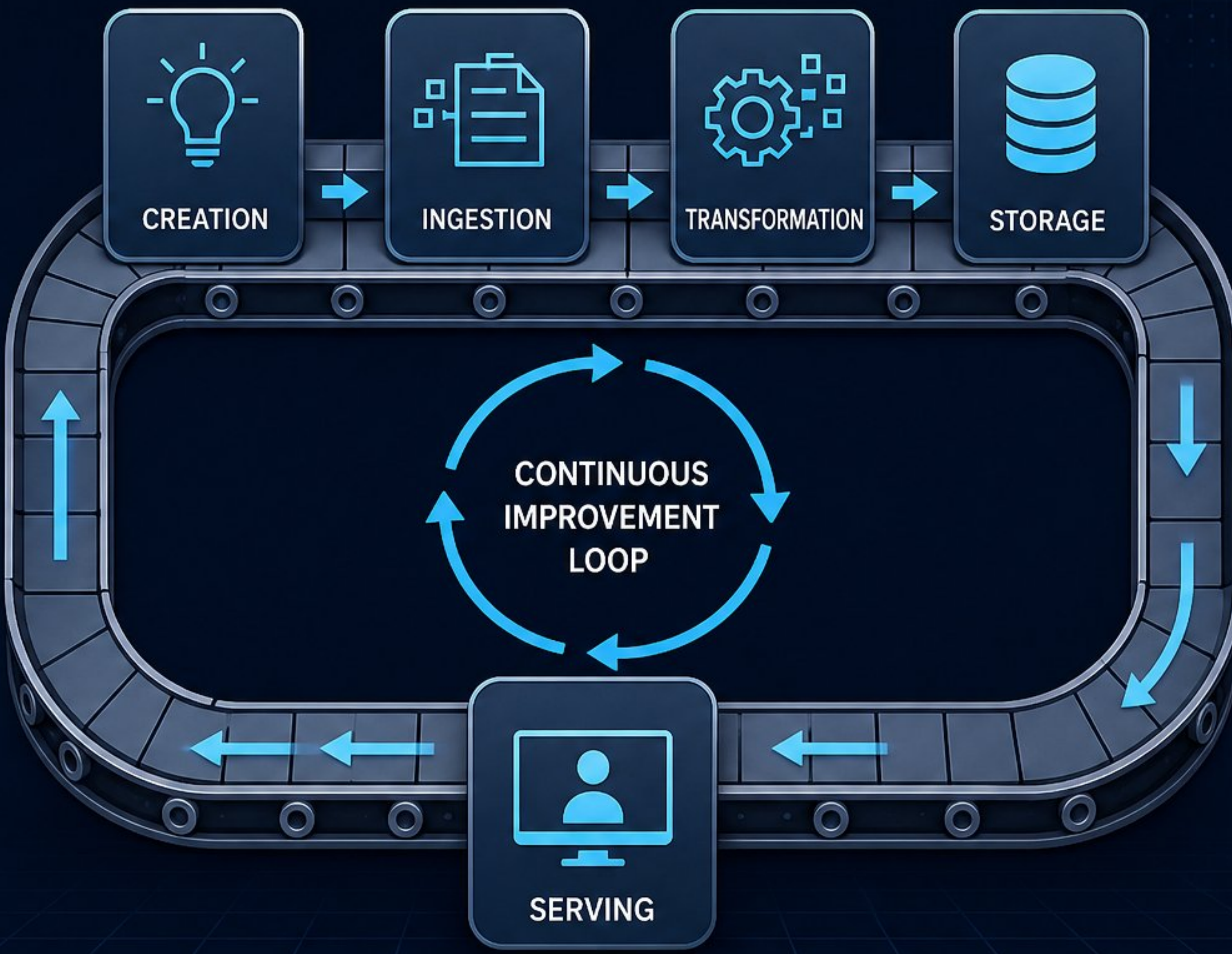
- Designing, building, and maintaining systems that collect, store, transform, and deliver data
- Closing the gap between raw, messy data and analysis-ready information
- The essential foundation connecting data sources, consumers, and business goals

Why Data Engineering Matters

- Messy, siloed, or unreliable data causes broken analytics and wrong decisions
- Trustworthy data accelerates decisions and powers product features
- Data engineering is the invisible foundation of sound decisions
- Raw data is messy; engineering turns it into reliable, analysis-ready data

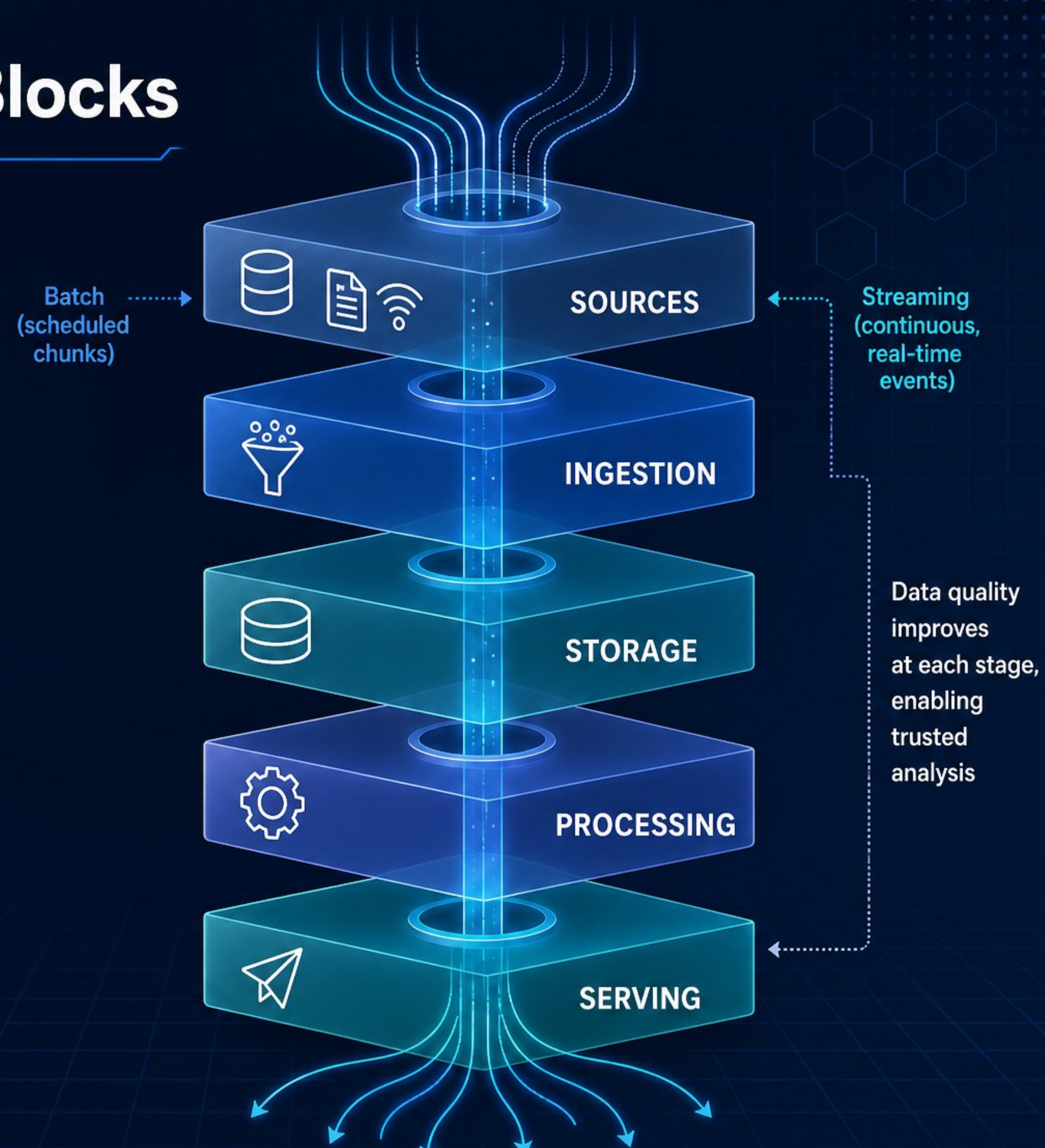
The Data Engineering Lifecycle

- Lifecycle spans creation, ingestion, transformation, storage, and serving
- Continuous improvement loop, not a one-time project
- Data engineers build, automate, and maintain reliable pipelines
- The role frames how data flows from source to consumption

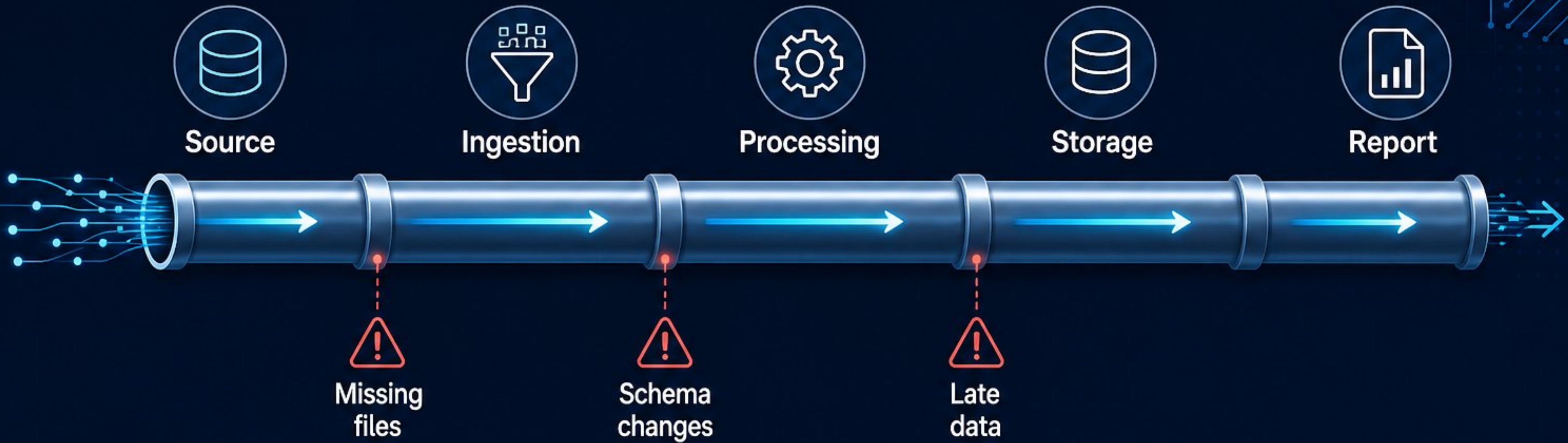


Core Building Blocks

- Five blocks: sources, ingestion, storage, processing, serving
- Batch: scheduled chunks; streaming: continuous, real-time events
- Raw data becomes analysis-ready through cleaning, transforming, and modeling
- Data quality improves at each stage, enabling trusted analysis



A Simple End-to-End Pipeline



- Pipeline moves data from source to report
- Common failures: missing files, schema changes, late data
- Delays happen at every stage, especially ingestion
- Scheduling ensures data arrives on time
- Monitoring and error handling catch failures early

Data Storage and Formats

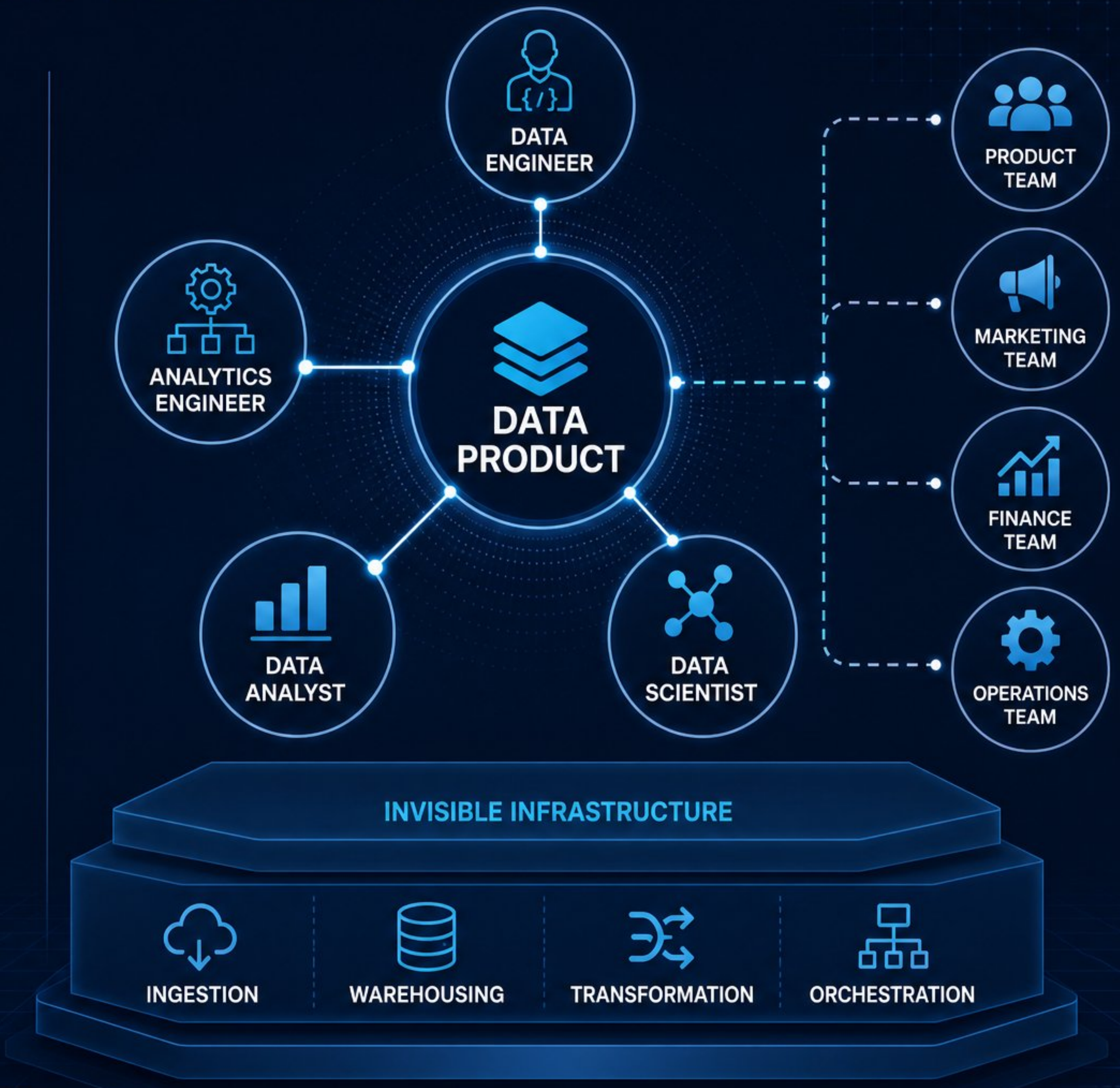
- Databases for operations; warehouses for analytics
- Data lakes store raw data in any format
- Data: structured, semi-structured, unstructured
- Formats matter: CSV, JSON, Parquet trade-offs
- Columnar formats like Parquet speed up queries

Data Quality and Trust

- Trustworthy data is accurate, complete, and fresh
- Common issues: missing values, duplicates, data drift, stale data
- Good quality data powers confident decisions and reliable reports
- Six core dimensions: completeness, uniqueness, timeliness, validity, accuracy, consistency
- Automated checks and clear ownership improve reliability

Roles and Team Workflows

- Core roles: data engineer, analytics engineer, data analyst, data scientist
- Data engineers: build and maintain pipelines, warehouses, and systems
- Analytics engineers: own transformation logic and semantic layers
- Data analysts: consume data to build reports and dashboards
- Data scientists: model data for predictions and experiments
- Collaboration: product, marketing, finance, and operations teams
- Tool categories: ingestion, warehousing, transformation, orchestration



Recognizing Data Engineering in Practice

- Dashboards, recommendations, and fraud alerts all sit on engineered pipelines
- Ask: What is the source? How fresh is it? Who owns it?
- Look for scheduled refreshes, documented transformations, and quality checks
- Start safely: query public datasets, use dbt locally, and explore via notebooks
- The invisible pipeline is why numbers match across teams

Key Takeaways and Next Steps

- From raw data to reliable insight, built on SQL and Python
- Start with SQL, then Python; build a simple pipeline that runs
- Focus on one cloud, one orchestrator, and one warehouse
- Paths into the field: from analytics, BI, or engineering roles
- Portfolio projects matter more than degrees or certifications

PersonWise.

PersonWise • AI digital-human course platform

Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/2b96d25d/public